Originalarbeit

Wie fair testet der WÜRT 1 die Rechtschreibleistungen von mehrsprachigen Kindern?

Eine Überprüfung mittels Differential Item Functioning

Janin Brandenburg[1,2], Sina S. Huschka[2,3], Linda Visser[2,4], Friederike Cartschau[5] und Ariane von Goldammer[5]

[1]Fachgebiet Partizipation bei Beeinträchtigungen des Lernens, Fakultät Rehabilitationswissenschaften, Technische Universität Dortmund, Deutschland
[2]Individual Development and Adaptive Education of Children at Risk (IDeA Zentrum), Frankfurt (Main), Deutschland
[3]DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformationnstitut, Frankfurt (Main), Deutschland
[4]Department of Developmental Psychopathology, Behavioural Science Institute, Radboud University Nijmegen, Niederlande
[5]Institut für Psychologie, Fachbereich: Erziehungs- und Sozialwissenschaften, Universität Hildesheim, Deutschland

Zusammenfassung: Bei der Diskussion über Bildungsgerechtigkeit für mehrsprachige Kinder steht u. a. die Testfairness standardisierter Schulleistungstests im Fokus. So sollte eine Leistungsdiagnostik dieser Kinder auf Tests zurückgreifen, deren Fairness für Kinder mit nichtdeutscher Muttersprache empirisch abgesichert ist, denn nur so lassen sich die Testergebnisse in gleicher Weise interpretieren wie bei Kindern, die einsprachig deutsch aufwachsen. Ein Ziel bestand daher darin, die Testfairness des Würzburger Rechtschreibtests für 1. und 2. Klassen (WÜRT 1 – 2) zu überprüfen. Außerdem wurde analysiert, ob mehrsprachige Kinder die gleichen Fehlerschwerpunkte in der Rechtschreibung aufweisen wie einsprachige Kinder. Es nahmen 146 einsprachige und 107 mehrsprachige Kinder am Ende der ersten Klasse teil. Analysen zum Differential Item Functioning zeigten nur bei einem der insgesamt 36 Items Hinweise auf eine systematische Benachteiligung mehrsprachiger Kinder. Mit dem WÜRT 1 liegt somit ein faires Testverfahren vor, dessen Einsatz auch bei mehrsprachigen Kindern empfohlen werden kann. Die ein- und die mehrsprachigen Kinder unterschieden sich in der Fehleranzahl, nicht aber ihren qualitativen Fehlerprofilen. Dies legt nahe, dass eine Rechtschreibförderung für mehrsprachige Kinder an den gleichen Schwerpunkten ansetzen kann wie bei einsprachigen Kindern.

Schlüsselwörter: Testfairness, Mehrsprachigkeit, Rechtschreibung, Differential Item Functioning (DIF), WÜRT 1 – 2

Does the WÜRT 1 Fairly Assess the Spelling Performance of Multilingual Children? A Differential Item Functioning Analysis

Abstract: We need standardized instruments with empirically supported fairness for the psychoeducational assessment of the school performance of multilingual children. Otherwise, it remains unknown whether we can validly compare the test results of multilingual children to those of monolingual children. Therefore, this study evaluated whether items of the Würzburger spelling test for 1st and 2nd grade (WÜRT 1 – 2) showed differential item functioning for multilingual (n = 107) and monolingual (n = 146) first-graders in Germany. Further, it examined whether the two groups make the same spelling errors. Results showed that only one of the 36 items systematically disadvantages multilingual children. Thus, the WÜRT 1 is a fair test that can be recommended for assessing multilingual first-graders. Group differences emerged in the number of errors but not in the qualitative error profiles. Spelling training for multilingual children can thus focus on the same topics as for monolingual children.

Keywords: test fairness, multilingualism, spelling, differential item functioning (DIF), WÜRT 1 – 2

Nach neusten Schätzungen ist Deutschland mittlerweile das zweitstärkste Einwanderungsland innerhalb der OECD (OECD, 2018). Diese Entwicklung spiegelt sich auch in den deutschen Schulen wider: Kinder mit Migrationshintergrund bilden eine zunehmend wachsende Subpopulation und machen gegenwärtig einen Anteil von 34 – 37 % der 6- bis 15-Jährigen aus (Autorengruppe Bildungsberichterstattung, 2018). Im Zuge der Debatte um Bildungsgerechtigkeit ist der schulische Wissens- und Kompetenzerwerb dieser Schüler_innengruppe immer wieder Gegen-

stand der schulpolitischen und wissenschaftlichen Diskussion, denn Kinder mit Migrationshintergrund zeigen im Vergleich zu Kindern ohne Migrationshintergrund oftmals Kompetenznachteile in der Schriftsprache (Blatt, Prosch & Lorenz, 2016; May, 2006; Schwippert, Wendt & Tarelli, 2012) sowie in Mathematik und den Naturwissenschaften (Gebhardt, Rauch, Mang, Sälzer & Stanat, 2013; Tarelli, Schwippert & Stubbe, 2012).

Eine entscheidende Determinante für den Bildungserfolg von Kindern mit Migrationshintergrund stellt dabei die Kompetenz in der Unterrichtssprache Deutsch dar. Denn viele der Kinder sprechen Deutsch als Zweitsprache oder erwerben Deutsch neben einer oder mehreren weiteren Erstsprache(n) (Stanat & Christensen, 2006). So lässt sich ein beachtlicher Anteil der Leistungsnachteile von Kindern mit Migrationshintergrund darauf zurückführen, ob und in welchem Umfang die Kinder auch in der Familie und in der Freizeit Deutsch sprechen und demnach auch außerhalb der Schule Lerngelegenheiten für den Erwerb der Unterrichtssprache haben (vgl. Tiedemann & Billmann-Mahecha, 2007).

Neben ihrer Bedeutung für den Wissenserwerb wird die Relevanz sprachlicher Kompetenzen aber auch beim standardisierten Testen zunehmend diskutiert, denn die Diagnostik schulischer Fertigkeiten von Kindern mit nichtdeutscher Muttersprache stellt die pädagogisch-psychologische Praxis vor besondere Herausforderungen: Um ein schwaches Testergebnis tatsächlich als Ausdruck eines zugrunde liegenden Defizites im domänenspezifischen Wissen interpretieren zu können, muss sich das verwendete Testverfahren zur Leistungsbeurteilung bei mehrsprachigen Kindern ebenso gut eignen wie für Kinder, die einsprachig mit Deutsch aufwachsen. Es sollte im Sinne der Testfairness also sichergestellt werden, dass mehrsprachige Kinder nicht systematisch benachteiligt werden. Erste Forschungsarbeiten in Deutschland haben diese Frage für ausgewählte standardisierte Testverfahren zur Erfassung der Intelligenz (WNV; Daseking et al., 2015) oder des Lesens (ELFE II; Lenhard & Lenhard, 2017) bereits beantwortet. Für das Rechtschreiben liegen hingegen noch keine Befunde vor. Dies verwundert, gibt es doch Hinweise darauf, dass Kinder, die mehrsprachig aufwachsen, im Durchschnitt geringere Rechtschreibkompetenzen im Deutschen aufweisen als solche, die einsprachig aufwachsen (Blatt et al., 2016; May, 2006; Schründer-Lenzen & Merkens, 2006), und daher ein erhöhtes Risiko für die Entwicklung von Rechtschreibschwierigkeiten zeigen. Die Frage nach einer fairen Rechtschreibdiagnostik, bei der die zugrunde liegenden Kompetenzen der Kinder nicht durch die besondere Spracherwerbsbedingung verzerrt werden, ist für die pädagogisch-psychologische Praxis daher höchst relevant. Das Ziel der vorliegenden Studie bestand daher darin zu untersuchen, ob der *Würzburger Rechtschreibtest für die erste und zweite Klasse* (WÜRT 1–2; Trolldenier, 2014) Testfairness für mehrsprachige Schüler_innen aufweist.

Testfairness

Testfairness liegt vor, wenn ein Test so konstruiert ist, dass er keine Personen entsprechend ihrer Zugehörigkeit zu bestimmten ethnischen oder (sozio-)kulturellen Gruppen systematisch benachteiligt und andere bevorzugt (siehe Hartig, Frey & Jude, 2012). Ein Test ist also dann fair, wenn bei seiner Anwendung für keine der damit getesteten Subgruppen eine systematische Über- oder Unterschätzung hinsichtlich der zugrunde liegenden Leistung entsteht (Cleary, 1968; siehe auch Lenhard & Lenhard, 2017). Ausschlaggebend ist dabei nicht, ob es tatsächlich Leistungsunterschiede zwischen den Gruppen gibt, sondern vielmehr, ob ein Verfahren zu subgruppenspezifischen Fehleinschätzungen der zugrunde liegenden Fähigkeiten führt.

Bei der standardisierten Leistungsdiagnostik von Kindern, deren Muttersprache nicht die Unterrichtssprache ist, kann nach García und Pearson (1994) die Testfairness vor allem durch drei Arten von Konstruktionsfehlern verletzt werden:

1. Ein *Linguistic Bias* liegt vor, wenn ein standardisierter Leistungstest unbeabsichtigt Anforderungen an die Sprachkompetenzen in der Unterrichtssprache stellt, obwohl diese Kompetenz nicht mit dem Test erfasst werden soll. Kinder mit unzureichenden Sprachkenntnissen laufen dann aufgrund sprachlicher Barrieren Gefahr, die Testaufgaben durch Defizite im deutschen Lese- oder Hörverstehen falsch oder unvollständig zu bearbeiten und ihr Wissen nicht entsprechend demonstrieren zu können.
2. Ein *Content Bias* besteht, wenn die Lösungsgüte in einem Leistungstest auf unbeabsichtigte Weise durch das kulturelle Wissen über das Aufnahmeland beeinflusst wird bzw. davon abhängt, wie stark eine kulturelle Integration im Aufnahmeland stattgefunden hat. Dies kann beispielsweise der Fall sein, wenn die in einem Schulleistungstest vorgegebenen Szenarien Lebenssituationen widerspiegeln, die typisch für die Kultur des Aufnahmelandes sind, aber nicht für die des Herkunftslandes.
3. Ein *Norming Bias* liegt schließlich vor, wenn die Testnormen die Spracherwerbssituation der Schulkinder nicht ausreichend berücksichtigen – etwa, wenn ein standardisierter Leistungstest (fast) ausschließlich an monolingual Deutschsprachigen normiert wurde, aber in der diagnostischen Praxis dennoch zur Leistungseinschätzung von Kindern mit anderer Muttersprache herangezogen wird.

Während sich die beiden erstgenannten Aspekte auf die sog. *prozedurale Testfairness* beziehen, wird der Norming Bias der *interpretationsbezogenen Testfairness* zugeordnet. Die interpretationsbezogene Testfairness stellt sicher, dass die aus den Testergebnissen resultierenden Entscheidungen über verschiedene Subgruppen hinweg gleichermaßen valide sind. Die prozedurale Testfairness hingegen beschäftigt sich mit der Frage, ob die Gestaltung der Testinstrumente oder die Testdurchführung bestimmte Personengruppen systematisch und unbeabsichtigt benachteiligt – oder anders ausgedrückt: inwiefern die Testaufgaben über verschiedene Gruppen hinweg gleich gut messen (siehe Schwabe & Gebauer, 2013). Die prozedurale Testfairness kann mit Analysen zum *Differential Item Functioning* (kurz: DIF) empirisch geprüft werden (siehe Holland & Wainer, 1993, für einen Überblick). Liegt bei einem oder mehreren Items DIF vor, hängt das Testergebnis von der Gruppenzugehörigkeit ab und spiegelt damit in mindestens einer der Gruppen die latente Fähigkeit nicht richtig wider. Ein Vergleich der Leistungen über die Gruppen hinweg wäre somit verzerrt. DIF-Analysen erlauben also eine Aussage darüber, ob die Testitems über verschiedene Gruppen hinweg (z. B. bei einversus mehrsprachigen Kindern) gleichermaßen zuverlässig messen. Der Fokus dieser Studie liegt auf der Untersuchung der prozeduralen Testfairness.

Der Erwerb der deutschen Rechtschreibung bei mehrsprachigen Kindern

Betrachtet man die spezifischen Herausforderungen, die mehrsprachige Kinder beim Erwerb der deutschen Rechtschreibung meistern müssen, so ist es durchaus plausibel zu vermuten, dass die Testfairness von Rechtschreibtests für diese Subgruppe nicht a priori angenommen werden kann. Denn neben den typischen Rechtschreibfehlern, die bei ihnen – genau wie auch bei einsprachig Deutsch aufwachsenden Kindern – vorkommen und die der falschen Anwendung deutscher Rechtschreibkonventionen entspringen, kommt es bei ihnen zusätzlich zu sog. *Interferenzfehlern.* Interferenzfehler liegen vor, wenn *phonologische Merkmale* oder *orthografische Regeln der Erstsprache* auf das Schreiben deutscher Wörter übertragen werden (Dahmen, 2012). Da mehrsprachige Erstklässler_innen meist noch keinen systematischen Rechtschreiberwerb in der nicht-deutschen Muttersprache erfahren haben, sind bei ihnen in erster Linie die phonologischen Aspekte relevant. D. h. Interferenzfehler kommen bei ihnen typischerweise dadurch zustande, dass ihre auditive Wahrnehmung und Aussprache deutscher Phoneme, Silben und Wörter durch die Hörgewohnheiten ihrer nicht-deutschen Muttersprache beeinflusst sind (Dahmen, 2012). Typische Interferenzfehler dieser Art betreffen nach Dahmen (2012; Dahmen & Weth, 2017) u. a. folgende Bereiche der deutschen Orthografie:

1. *Probleme bei der Phonemunterscheidung.* Da Sprachen sich in der Art, Anzahl und Qualität ihrer Phoneme unterscheiden, kann es bei mehrsprachigen Kindern zu Fehlern in der deutschen Rechtschreibung kommen, wenn es den Kindern (noch) schwer fällt, die verschiedenen Phoneme der deutschen Sprache richtig wahrzunehmen und zu unterscheiden – etwa weil entsprechende Laute in ihrer nicht-deutschen Muttersprache nicht vorkommen. Fehlende oder andere lautliche Kontraste in der Muttersprache können entsprechend zu Problemen im Hörverstehen diktierter deutscher Wörter führen, was sich in der Verschriftlichung entsprechend falsch niederschlägt.
2. *Konsonantencluster*. Im Deutschen sind im Gegensatz zu vielen anderen Sprachen sehr komplexe Silbenstrukturen durch Konsonantenhäufungen im Anlaut (z. B. „Straße“) oder Auslaut (z. B. „impfst“) möglich. Mehrsprachige Kinder, deren nicht-deutsche Muttersprache deutlich simplere Silbenstrukturen aufweist (z. B. Arabisch und Türkisch), haben folglich Schwierigkeiten, die komplexen Konsonantenhäufungen des Deutschen auszusprechen und entsprechend zu verschriftlichen. Typischerweise werden von den Kindern einzelne Konsonanten bei der Verschriftlichung weggelassen oder Sprossvokale eingefügt, um die Konsonantencluster „aufzubrechen“ (z. B. „beraun“ für „braun“).
3. *Vokallänge.* Im Deutschen ist die Vokallänge bedeutungsunterscheidend (z. B. Stadt versus Staat) und wird beispielsweise durch Dehnungs-h oder Doppelvokal verschriftlicht. Kinder, deren nicht-deutsche Muttersprache keine Unterscheidung in der Vokallänge kennt (z. B. Türkisch, Französisch, Russisch), hören und sprechen den Längenunterschied in deutschen Wörtern ggf. nicht, wodurch Fehler in der orthografischen Markierung der Vokallänge entstehen.

Diese Beispiele verdeutlichen, dass mehrsprachige Kinder mitunter Rechtschreibfehler zeigen, die durch ihre besondere Spracherwerbssituation zustande kommen. Dennoch scheinen erste Studien dafür zu sprechen, dass diese Fehlerarten oft nur einen kleinen Anteil an den Rechtschreibfehlern mehrsprachiger Kinder ausmachen und der weitaus größere Teil auf Fehler entfällt, die von ein- und mehrsprachigen Kindern gleichermaßen gemacht werden (siehe Ruppert & Hanulíková, im Druck).

Fragestellungen

Angesichts des nicht unerheblichen Anteils an mehrsprachigen Schulkindern in Deutschland ist anzunehmen, dass die Überprüfung des schulischen Kompetenzniveaus bei Kindern nicht-deutscher Muttersprache anhand standardisierter Leistungstests längst zum diagnostischen Alltag in Beratungsstellen und sozialpädiatrischen Zentren wie auch in den Schulen selbst geworden ist. Im Rahmen einer solchen Leistungsdiagnostik sollte idealerweise ein Testverfahren herangezogen werden, dessen Fairness für diese Gruppe empirisch geprüft wurde. Für viele standardisierte Schulleistungstests in Deutschland steht eine solche Überprüfung jedoch noch aus. Vor diesem Hintergrund hatte die vorliegende Studie das Ziel, den *Würzburger Rechtschreibtest für die erste und zweite Klasse* (WÜRT 1–2; Trolldenier, 2014) hinsichtlich seiner Tauglichkeit für mehrsprachige Erstklässler_innen zu überprüfen. Dabei wurden zwei Forschungsfragen getestet:

1. Zeigen die Items des WÜRT 1 prozedurale Testfairness hinsichtlich des Merkmals Mehrsprachigkeit?
2. Unterscheiden sich ein- und mehrsprachige Schulkinder hinsichtlich ihrer Fehlerprofile im WÜRT 1?

Methode

Stichprobe

Die Daten stammen aus der Längsschnittstudie TRIO, bei der Datenerhebungen sowohl im Kindergarten als auch in der Grundschule stattfanden. TRIO war eine Studie im Rahmen der Bund-Länder-Initiative *Bildung durch Sprache und Schrift* (BiSS). Die Rekrutierung der 305 Kinder erfolgte in 24 deutschsprachigen Kindergärten im südwestlichen Rhein-Main-Gebiet. Der Erstkontakt mit den Einrichtungen erfolgte über die BiSS-Landeskoordination in Hessen, worüber öffentliche Grundschulen im südwestlichen Rhein-Main-Gebiet über die geplante Studie informiert und gebeten wurden, mit ihren jeweiligen, über die hessischen Vorlaufkurse assoziierten Kindertagesstätten ein (unverbindliches) Teilnahmeinteresse zu erörtern. Interessierte Tandems, bestehend aus Grundschule und Kita, wurden zu einer Informationsveranstaltung über die geplante Studie eingeladen und hatten anschließend die Möglichkeit, sich verbindlich zur Teilnahme anzumelden. Die teilnehmenden Kitas erhielten Einladungsschreiben zu der geplanten Studie, mit der Bitte diese an diejenigen Erziehungsberechtigten zu verteilen, deren Kind das letzte Kindergartenjahr besuchte. Über die Einladungsschreiben wurden die Familien über die Studie informiert und konnten ihr Kind für die Teilnahme anmelden. Von 16 der 24 teilnehmenden Kindergärten liegen Informationen zur Rücklaufquote der eingeladenen Familien vor. Demnach lag der Anteil der teilnehmenden Kinder gemessen an der Gesamtzahl der Kindergartenkohorte im Mittel bei 75 % (*SD* = 14 %, Range = 43–100 %).

Die Rechtschreibleistung der Kinder wurde am Ende der ersten Klasse erfasst. Bis zu diesem Zeitpunkt waren 37 Kinder (12 % der Ursprungsstichprobe) aus verschiedenen Gründen (z. B. Wohnortwechsel, Rückstufung) aus der Studie ausgeschieden. Des Weiteren konnten die Testhefte von 15 Kindern (5 %) nicht ausgewertet werden, weil die Kinder beispielsweise kein einziges Item geschrieben hatten oder lediglich die im Testheft abgedruckten Wörter „abgeschrieben“ hatten, anstatt die diktierten Wörter zu verschriftlichen. Somit flossen die Daten von 253 Kindern aus 31 Schulen in die vorliegende Analyse ein: 146 Kinder (58 %; darunter 80 Jungen; Alter in Jahren: M = 7.40, SD = 0.29) wuchsen mit Deutsch als alleiniger Muttersprache auf, während die restlichen 107 Kinder (42 %; darunter 55 Jungen; Alter in Jahren: M = 7.42, SD = 0.31) mehrsprachig aufwuchsen. Die Einteilung in diese beiden Spracherwerbsgruppen erfolgte anhand von Elternangaben zur Muttersprache des Kindes (d. h. die zuerst gelernte Sprache). Innerhalb der mehrsprachigen Gruppe sprachen 59 Kinder neben Deutsch noch mindestens eine weitere Muttersprache (bilinguale Kinder), während 48 Kinder eine andere Muttersprache sprachen und erst nach dem 2. Lebensjahr mit dem Deutschen in Kontakt gekommen waren (Zweitsprachlernende). Bilingual aufwachsende Kinder und Kinder mit Deutsch als Zweitsprache wurden in der vorliegenden Studie zu der gemeinsamen Gruppe der mehrsprachigen Kinder zusammengefasst, weil die beiden Teilgruppen zu klein gewesen wären, um die Analysen der Testfairness getrennt durchzuführen. Explorativ wurden jedoch zusätzlich die Analyse von Gruppenunterschieden in der Fehleranzahl und die Analyse der Fehlerprofile mit einer Einteilung in drei Gruppen (einsprachig – zweitsprachlernend – bilingual) durchgeführt. Neben Deutsch wurden in der Gruppe der mehrsprachigen Kinder insgesamt 30 verschiedene Sprachen gesprochen, wobei Türkisch und Arabisch mit 22 bzw. 14 % am häufigsten vertreten waren.

Die Altersverteilung und das Geschlechterverhältnis waren in den beiden Gruppen annähernd identisch und unterschieden sich nicht statistisch signifikant zwischen den Gruppen, $t(251) < -1$, $p = .557$, $d_{\text{Cohen}} = -0.08$ für das Alter und $\chi^2(1) < 1$, $p = .593$ für das Geschlecht. Auch wurde geprüft, ob die beiden Gruppen unterschiedlich gute Leistungen in den vier Diktaten des WÜRT 1 erzielten. Während die einsprachigen Kinder sowohl im ersten Diktat, $t(251) = 2.00$, $p = .046$, $d_{\text{Cohen}} = 0.26$, als auch im dritten Diktat, $t(246.77) = 4.72$, $p < .001$, $d_{\text{Cohen}} = 0.59$, im Mittel bedeutsam besser abschnitten als die mehrsprachi-

gen Kinder, war dies beim zweiten, $t(251) = 1.16$, $p = .247$, $d_{Cohen} = 0.15$, und vierten Diktat, $t(251) = 1.39$, $p = .167$, $d_{Cohen} = 0.18$, nicht der Fall. Bezogen auf das Gesamtergebnis über alle vier Diktate hinweg ergab sich ein signifikanter Gruppenunterschied, $t(251) = 2.67$, $p = .008$, $d_{Cohen} = 0.34$.

Die Stichprobe verteilte sich auf 31 Schulen, wobei im Mittel acht Kinder auf eine Schule entfielen ($SD = 10.84$). In etwas über einem Drittel der Fälle (36 %) war es so, dass lediglich ein Kind die jeweilige Schule besuchte; am zweithäufigsten (mit 19 % der Fälle) kamen zwei teilnehmende Kinder pro Schule vor. D. h. in 55 % der Fälle war mit lediglich ein bis zwei Schüler_innen pro Schule das Clustering der Stichprobe sehr gering. Die Intraklassenkorrelation (ICC) wurde bestimmt, um zu prüfen, wie viel Varianz in der Rechtschreibleistung auf Unterschiede zwischen den Schulen zurückführbar ist. Mit einem Wert von .03 war die ICC in der vorliegenden Studie vernachlässigbar gering.

Würzburger Rechtschreibtest für 1. und 2. Klassen (WÜRT 1 – 2)

Die Rechtschreibleistung wurde mit der Diktatform für die erste Klasse, dem WÜRT 1 (Trolldenier, 2014), erfasst. Im Rahmen von Lückentextdiktaten werden den Kindern 36 Items diktiert, die in vier thematische Geschichten eingebettet sind. Die Items lassen sich drei Kategorien zuordnen: Die *Mitsprechwörter* umfassen 13 lauttreue Wörter, die genauso geschrieben werden, wie man sie spricht. Die Gruppe der *Nachdenkwörter* setzt sich aus zehn Wörtern mit phonologischen Regelhaftigkeiten der deutschen Schriftsprache zusammen (z. B. die Schreibung von „ei" für den Laut [ai]). Die 13 *Merkwörter* schließlich prüfen die Anwendung von Rechtschreibregeln, welche nicht erschlossen, sondern nur erlernt werden können (z. B. Dehnungs-h oder Konsonantenverdopplung).

Jedes falsch geschriebene oder ausgelassene Wort zählt als Fehler. Neben dieser quantitativen Auswertung ist entsprechend des Testmanuals eine qualitative Auswertung der Rechtschreibfehler möglich, welche in der vorliegenden Studie von geschulten studentischen Mitarbeitenden durchgeführt wurde. Im WÜRT 1 werden neun phänomenologisch-deskriptive Fehlerkategorien unterschieden, wobei innerhalb eines Wortes mehrere Fehler kodiert werden können: Auslassung eines Lautes (A), Hinzufügen eines Lautes (H), Lautverwechslung (LV), Lautgetreu aber falsch (LF), Umstellung von Buchstaben innerhalb eines Wortes (U), Verstöße gegen die Groß- und Kleinschreibung (GK), Starke Wortbildveränderung (SW), Auslassen eines ganzen Wortes (AW) und Restfehler (R).

Der WÜRT 1 ist für Kinder am Ende des ersten Schuljahres bzw. am Anfang der zweiten Klasse normiert. Die Normierungsstichprobe umfasst 2 742 Kinder (46 % Jungen) aus sieben Bundesländern, wobei Hessen – das Bundesland, in dem die vorliegende Studie durchgeführt wurde – nicht dabei war. Für 82 % der Normierungsstichprobe standen Angaben der Lehrkraft zur Muttersprache der Kinder (d. h. Erstsprache des Kindes) zur Verfügung. Demnach hatten 83 % der Kinder die Muttersprache Deutsch; bei den anderen Muttersprachen waren Türkisch (7 %) und Russisch (4 %) am häufigsten vertreten. Die Reliabilität des WÜRT 1 ist im Testmanual mit einem Cronbachs Alpha von .91 angegeben und betrug in der vorliegenden Stichprobe $\alpha = .89$. In Bezug auf die Rechtschreibleistung der vorliegenden Studie zeigt sich, dass sowohl die einsprachige Gruppe mit einem mittleren T-Wert von 45.49 ($SD = 8.39$) als auch die mehrsprachige Gruppe mit einem mittleren T-Wert von 42.63 ($SD = 8.63$) gemessen an der Normierungsstichprobe des WÜRT 1 eher niedrig abschnitten: Der Leistungsunterschied zum Erwartungswert von T = 50 erwies sich im Einstichproben-*t*-Test in beiden Fällen als statistisch signifikant, $t_{einsprachig}(145) = -6.50$, $p < .001$, $d_{Cohen} = 0.45$; $t_{mehrsprachig}(106) = -8.63$, $p < .001$, $d_{Cohen} = 0.74$. Im Vergleich zur Normierungsstichprobe erzielten die einsprachigen Kinder in dieser Studie im Mittel 4 Rohwert-Punkte weniger und die mehrsprachigen Kinder sogar 8 Rohwert-Punkte weniger.

Statistische Auswertung

Zur Überprüfung von DIF wurde ein „*latent mixture modeling*"-Ansatz mit vordefinierten Klassen (siehe Raykov, Dimitrov, Marcoulides, Li & Menold, 2018) gewählt. Die Analysen wurden in Mplus Version 8.4 (Muthén & Muthén, 1998 – 2020) durchgeführt. Die zwei latenten Klassen (ein- versus mehrsprachige Kinder) wurden mit der „Knownclass"-Option spezifiziert. Bei diesem Ansatz werden Modelle der *Item Response Theorie* im theoretischen Kontext von Strukturgleichungsmodellen analysiert. Weil mit binären Indikatoren (hier: die Items im WÜRT, die entweder falsch oder korrekt gelöst wurden) gearbeitet wird und die Varianz des unidimensionalen latenten Konstruktes (hier: die Rechtschreibleistung) auf 1 gesetzt wird, ist das resultierende Modell im Sinne eines 2PL-Modells der IRT interpretierbar. Die Faktorladungen im Modell spiegeln demnach die Itemtrennschärfen wider und die Thresholds die Itemschwierigkeiten des Rechtschreibtests. Durch die Spezifizierung der vordefinierten Klassen folgt dieses Modell zusätzlich dem Mehrgruppen-Ansatz von Strukturgleichungsmodellen, wodurch die Modellparameter (d. h. Faktorladungen und Thresholds) getrennt

für die beiden Gruppen berechnet werden. Die entsprechende Syntax ist im elektronischen Supplement (ESM) 1 einzusehen (siehe auch Raykov et al., 2018). Konkret schlagen Raykov et al. (2018) ein vierstufiges Testverfahren vor:

1. Berechnung gruppengetrennter konfirmatorischer Faktorenanalysen zur Prüfung des Messmodells. Ein guter Modellfit wurde entsprechend der Empfehlungen von Hu und Bentler (1999) definiert über (a) eine nicht signifikante χ^2-Statistik und ein Verhältnis von $\chi^2/df > 2$, (b) einen *Root Mean Square Error of Approximation* (RMSEA) ≤ .06, (c) einen *Comparative Fit Index* (CFI) ≥ .95 sowie ein *Weighted Root Mean Square Residual* (WRMR) < .90.
2. Spezifizierung des oben beschriebenen *mixture*-Modells mit gleichgesetzten Faktorladungen und Thresholds für die beiden Gruppen (Baselinemodell M0).
3. Sukzessive Freisetzung der Faktorladungen und Thresholds für die *n*-Items eines Diktates, wodurch 2**n* verschiedene Testmodelle M1 entstehen, welche mithilfe von χ^2-Modellvergleichen mit dem Baselinemodell verglichen werden.
4. Mithilfe der *Benjamini-Hochberg-(B-H)-Methode* (Benjamini & Hochberg, 1995) wird das multiple Testen in Schritt 3 kontrolliert. Dabei wird für jeden *p*-Wert der individuelle kritische B-H Wert berechnet, der unterschritten sein muss, um die statistische Signifikanz abzusichern. Wird keiner dieser korrigierten *p*-Werte unterschritten, gibt es kein DIF.

Im zweiten Auswertungsschritt wurde mit einer manifesten Profilanalyse in SPSS geprüft, ob sich bei den ein- und mehrsprachigen Kindern die Fehlerschwerpunkte in der Rechtschreibung voneinander unterscheiden. Dazu wurden die mittleren Häufigkeiten der Fehlerkategorien mithilfe einer zweifaktoriellen Messwiederholungs-ANOVA analysiert, wobei die Fehler den multivariaten Messwiederholungsfaktor bildeten und der Spracherwerbstyp als Zwischensubjektfaktor diente. Von den neun Fehlerkategorien des WÜRT 1 wurden dabei alle Kategorien mit Ausnahme der „Restfehler" in die Analyse einbezogen, da diese Kategorie laut Manual aufgrund ihrer geringen Vorkommenshäufigkeit praktisch nicht bedeutsam ist. Auch in der vorliegenden Studie wurde diese Kategorie insgesamt nur vier Mal kodiert.

In einer zusätzlichen Analyse wurden schließlich auch die Fehlerprofile im Rahmen eines Drei-Gruppen-Modells geprüft, um zu explorieren, ob sich die einsprachigen, die zweisprachigen und die bilingualen Kinder in ihren Rechtschreibfehlern voneinander unterscheiden.

Ergebnisse

Überprüfung von DIF

Diktat #1 „*Auf der Wiese*". Für das erste Diktat ergab sich für die einsprachige Gruppe ein guter, $\chi^2(27) = 43.35$, $p = .024$; $\chi^2/df > 2$; RMSEA = .06, 90 % KI [.02, .09]; CFI = .96, WRMR = .89, und für die mehrsprachige Gruppe ein sehr guter Modellfit, $\chi^2(27) = 29.96$, $p = .316$; $\chi^2/df > 2$; RMSEA = .03, 90 % KI [.00, .08]; CFI = .99, WRMR = .65. Anschließend wurde das Baselinemodell M0 spezifiziert und daran anknüpfend die neun Faktorladungen und Thresholds in jeweils verschiedenen Modellen (W1(1) – W1(18)) über die Gruppen freigesetzt. Ein Vergleich dieser 18 Modelle mit dem Baselinemodell ergab keine signifikanten Unterschiede zwischen den Modellen. Die Annahme invarianter Itemschwierigkeiten und Trennschärfen bei gleichen zugrunde liegenden Rechtschreibfähigkeiten von ein- bzw. mehrsprachigen Kindern kann somit angenommen werden. Es liegt also kein DIF vor. Die Ergebnisse dieser Analyse sind im ESM 2 einzusehen.

Diktat #2 „*Im Krankenhaus*". Beim zweiten Diktat stellte sich im Rahmen der konfirmatorischen Faktorenanalyse für beide Gruppen heraus, dass das vierte Item – das Wort „Rücken" – mit standardisierten Faktorladungen von .24 (mehrsprachige Gruppe) bzw. von .06 (einsprachige Gruppe) nicht signifikant durch den zugrunde liegenden Rechtschreibfaktor abgebildet wurde. Auch zeigte sich bei Inspektion dieses Items, dass es nur von etwa 2 % der Kinder beider Gruppen richtig geschrieben wurde. Dies deckt sich mit Angaben aus dem Testmanual, bei dem das Wort „*Rücken*" die höchste Itemschwierigkeit innerhalb des gesamten WÜRT 1 aufweist. Bei den folgenden DIF-Analysen wurde das Item aufgrund seiner geringen Bearbeitungsgüte sowie der nicht signifikanten, Faktorladungen ausgeschlossen. Das resultierende Messmodell aus den übrigen Items erbrachte für beide Gruppen einen sehr guten Modellfit, einsprachige Kinder: $\chi^2(27) = 23.48$, $p = .659$; $\chi^2/df > 2$; RMSEA = .00, 90 % KI [.00, .05]; CFI = 1.00, WRMR = .59; mehrsprachige Kinder: $\chi^2(27) = 37.53$, $p = .085$; $\chi^2/df > 2$; RMSEA = .06, 90 % KI [.00, .10]; CFI = .95, WRMR = .77.

Die sukzessive Freisetzung der Thresholds bzw. Faktorladungen und deren Vergleich mit dem Baselinemodell ergab, dass der Threshold von Item 2 – d. h. die Itemschwierigkeit des Wortes „seinen" – nicht zwischen den Gruppen gleichgesetzt werden konnte, demnach also DIF vorlag (siehe ESM 3). Während die einsprachigen Kinder bereits bei einer zugrunde liegenden Rechtschreibfähigkeit von $\theta = -1.07$ das Wort zu 50 % richtig verschriftlichen konnten, benötigten die mehrsprachigen Kinder eine la-

tente Rechtschreibfähigkeit von θ = -0.36, um für dieses Wort eine mittlere Lösungswahrscheinlichkeit zu erzielen. Nach Absicherung dieses Befundes gegen den statistischen Zufall unter Hinzunahme der B-H-Methode blieb dieser Befund jedoch nicht bestehen (Threshold von Item 2: $p_{\tau 02}$ = .027; errechneter B-H Cut-off Wert für Item 2: $p_{B\text{-}H}$ = .003; $p_{\tau 02} > p_{B\text{-}H}$).

Diktat #3 „*Im Park*". Für das dritte Diktat ergab sich für die einsprachige Gruppe ein guter Modellfit, $\chi^2(27)$ = 45.92, p = .013; $\chi^2/df > 2$; RMSEA = .07, 90 % KI [.03, .10]; CFI = .94, WRMR = .87, wobei jedoch erneut der χ^2-Test signifikant wurde und zusätzlich der RMSEA knapp den Cut-off von .06 verfehlte. Für die mehrsprachige Gruppe zeigte sich ein sehr guter Modellfit, $\chi^2(27)$ = 33.77, p = .173; $\chi^2/df > 2$; RMSEA = .05, 90 % KI [.00, .09]; CFI = .95, WRMR = .80. Die sukzessive Freisetzung der Thresholds bzw. Faktorladungen und deren Vergleich mit dem Baselinemodell ergab, dass der Threshold von Item 8 – d. h. die Itemschwierigkeit des Wortes „wünscht" – nicht zwischen den Gruppen gleichgesetzt werden konnte, demnach also DIF vorlag (siehe ESM 4). Während die einsprachigen Kinder bereits bei einer zugrunde liegenden Rechtschreibfähigkeit von θ = 1.18 das Wort zu 50 % richtig verschriftlichen konnten, benötigten die mehrsprachigen Kinder eine latente Rechtschreibfähigkeit von θ = 4.35, um für dieses Wort eine mittlere Lösungswahrscheinlichkeit zu erzielen. Auch unter Hinzunahme der B-H-Methode blieb dieser Befund bestehen (Threshold von Item 8: $p_{\tau 08} < .001$; errechneter B-H Cut-off Wert für Item 8: $p_{B\text{-}H}$ = .003; $p_{\tau 08} < p_{B\text{-}H}$). Dieses Item scheint mehrsprachige Kinder somit zu benachteiligen. Für alle anderen Thresholds und Faktorladungen kann die Invarianz über die Gruppen hingegen angenommen werden.

Diktat #4 „*Im Freien*". Für das vierte Diktat zeigte sich für beide Gruppen ein sehr guter Modellfit, einsprachige Kinder: $\chi^2(20)$ = 23.12, p = .283; $\chi^2/df > 2$; RMSEA = .03, 90 % KI [.00, .08]; CFI = .95, WRMR = .71; mehrsprachige Kinder: $\chi^2(20)$ = 14.12, p = .824; $\chi^2/df > 2$; RMSEA = .00, 90 % KI [.00, .05]; CFI = 1.00, WRMR = .54. Die sukzessive Freisetzung der Thresholds bzw. Faktorladungen und deren Vergleich mit dem Baselinemodell ergab, dass der Threshold von Item 4 – d. h. die Itemschwierigkeit des Wortes „Garten" – nicht zwischen den Gruppen gleichgesetzt werden konnte, demnach also DIF vorlag (siehe ESM 5). Hierbei war es erstaunlicherweise so, dass das Item den mehrsprachigen Kindern leichter fiel als den einsprachigen Kindern: Um für dieses Wort eine mittlere Lösungswahrscheinlichkeit zu erzielen, benötigten die einsprachigen Kinder eine latente Rechtschreibfähigkeit von θ = 1.03, während die mehrsprachigen Kinder bereits bei einer zugrunde liegenden Rechtschreibfähigkeit von θ = 0.54 das Wort zu 50 % richtig verschriftlichen konnten. Unter Hinzunahme der B-H-Methode blieb dieser Befund jedoch nicht bestehen (Threshold von Item 4: $p_{\tau 04}$ = .009; errechneter B-H Cut-off Wert für Item 4: $p_{B\text{-}H}$ = .003; $p_{\tau 04} > p_{B\text{-}H}$).

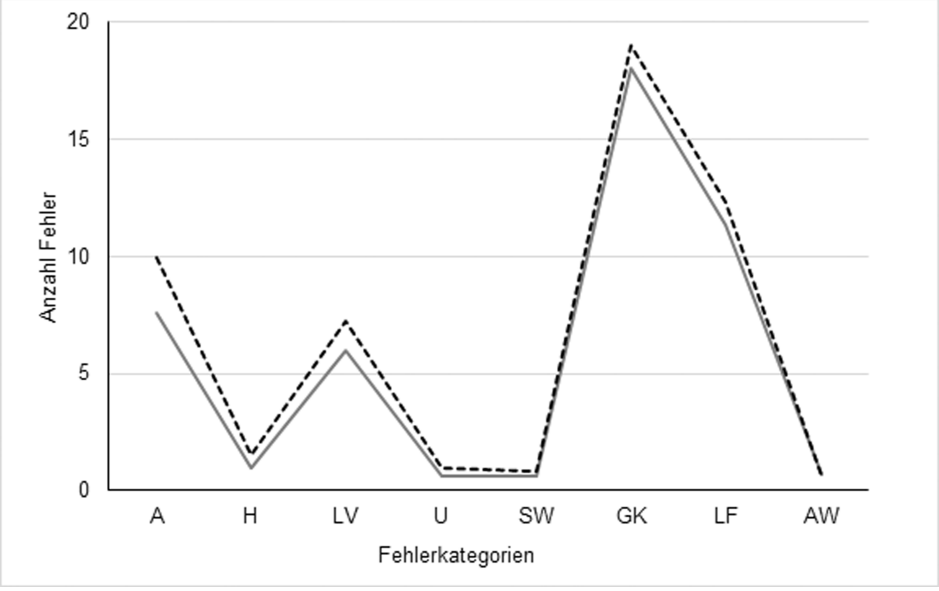
Profilanalyse der Rechtschreibfehler

Die Verteilung in den Fehlerkategorien ist in Abbildung 1 einzusehen. Da sich mit Ausnahme der Kategorien LV, GK und LF auffällige Werte entweder in der Schiefe (Werte > 3) oder in der Kurtosis (Werte > 8) in mindestens einer der beiden Gruppen zeigten, wurden die Analysen nicht nur mit den Rohwerten, sondern zusätzlich auch mit Logarithmus-transformierten Werten durchgeführt. Beide Analysevarianten ergaben jedoch dasselbe Ergebnismuster, weshalb im Folgenden lediglich die statistischen Kennwerte der nicht transformierten Rohwerte angegeben werden.

Dabei zeigte sich ein statistisch signifikanter Haupteffekt für den messwiederholten Faktor *Fehlerkategorie*, $F(3.57, 889.28)$ = 668.79, $p < .001$, η_p^2 = .73 (Ergebnisse unter Verwendung der Greenhouse-Geisser-Korrektur). Post-hoc Vergleiche mit Bonferroni-Korrektur ergaben, dass die meisten der 28 paarweisen Vergleichen statistisch signifikant waren. Lediglich die vier Fehlerkategorien, die am seltensten vorkamen (d. h. H, U, SW und AW) unterschieden sich in ihrer mittleren Auftretenshäufigkeit nicht bedeutsam voneinander. Wie auch in Abbildung 1 dargestellt, kamen Fehler in der GK sowie in den Kategorien LF und A in beiden Gruppen am häufigsten vor. Die Ergebnisse dieser Analyse sind in ESM 6 einzusehen. Zusätzlich ergab sich ein signifikanter – wenn auch schwacher – Haupteffekt für den Faktor *Spracherwerbstyp*, $F(1, 249)$ = 10.34, p = .001, η_p^2 = .04. Die mehrsprachigen Kinder machten im Mittel signifikant mehr Fehler als die einsprachigen Kinder. Die Interaktion zwischen den beiden Faktoren (Fehlerkategorie x Spracherwerbstyp) wurde statistisch nicht signifikant, $F(3.57, 889.28)$ = 2.24, p = .071, η_p^2 = .009, was sich grafisch in zwei Fehlerprofilen mit einem annähernd parallelen Verlauf darstellt.

Explorative Analysen mit drei Gruppen

Explorativ wurden anschließend die Gruppenunterschiede in den Rohwerten und in den Fehlerprofilen mit einer Einteilung in drei Gruppen (einsprachig – zweitsprachlernend – bilingual) inferenzstatistisch geprüft. Im Gegensatz zur vorherigen Analyse mit zwei Gruppen, bei der sich signifikante Gruppenunterschiede lediglich im ersten und dritten Diktat zeigten, wurde der Gruppenfaktor nun zusätzlich im zweiten Diktat statistisch signifikant, Diktat 1: $F(2, 250)$ = 4.48, p = .012, η_p^2 = .04; Diktat 2: $F(2, 250)$ = 3.89, p = .022, η_p^2 = .03; Diktat 3: $F(2, 250)$ = 11.28,



Anmerkungen: A = Auslassung eines Lautes, H = Hinzufügen eines Lautes, LV = Lautverwechslung, U = Umstellung von Buchstaben innerhalb eines Wortes, SW = Starke Wortbildveränderung, GK = Verstöße gegen die Groß- und Kleinschreibung, LF = Lautgetreu aber falsch, AW = Auslassen eines ganzen Wortes.

Abbildung 1. Fehlerprofile der einsprachigen Kinder (durchgezogene Linie) und mehrsprachigen Kinder (gestrichelte Linie) im WÜRT 1+ (dargestellt sind die durchschnittlichen Fehleranzahlen der beiden Gruppen).

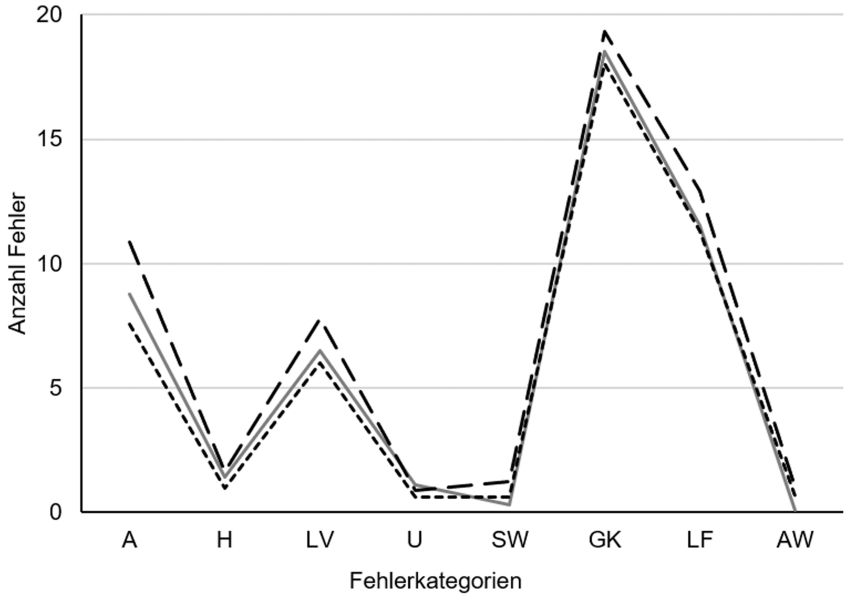
$p < .001$, $\eta_p^2 = .15$. Post-Hoc-Vergleiche mit Bonferroni-Korrektur ergaben, dass die Zweitsprachlernenden signifikant schwächere Ergebnisse in den drei Diktaten im Vergleich zu den einsprachigen Kindern erzielten, während die anderen paarweisen Vergleiche nicht signifikant wurden. Im vierten Diktat bestanden keine statistisch signifikanten Unterschiede zwischen den drei Gruppen, $F(2, 250) = 2.22$, $p = .110$, $\eta_p^2 = .017$. Bezogen auf das Gesamtergebnis über alle vier Diktate hinweg ergab sich ein statistisch signifikanter Gruppenunterschied, $F(2, 250) = 6.34$, $p = .002$, $\eta_p^2 = 0.05$. Im Post-Hoc-Vergleich wurde erneut lediglich der mittlere Rohwertunterschied zwischen den einsprachigen Kindern und den Zweitsprachlernenden statistisch signifikant.

Anschließend wurden die Fehlerprofile der drei Spracherwerbsgruppen varianzanalytisch miteinander verglichen. Eine ANOVA mit Messwiederholung (und Greenhouse-Geisser-Korrektur) erbrachte erneut einen bedeutsamen Haupteffekt für den messwiederholten Faktor Fehlerkategorie, $F(3.57, 884.71) = 553.48$, $p < .001$, $\eta_p^2 = .69$. Post-hoc Vergleiche mit Bonferroni-Korrektur zeigten die gleichen signifikanten Unterschiede in den mittleren Auftretenshäufigkeiten der Fehlerkategorien wie in den Analysen mit zwei Gruppen mit der Ausnahme, dass die mittlere Differenz zwischen den Fehlerkategorien H und AW nun signifikant wurde. Die Ergebnisse dieser Analyse sind in ESM 7 einzusehen. Außerdem wurde der Faktor Gruppe statistisch signifikant, $F(2, 248) = 8.10$, $p < .001$, $\eta_p^2 = .06$. Post-Hoc Vergleiche mit Bonferroni-Korrektur ergaben, dass die Zweitsprachlernenden im Mittel signifikant mehr Fehler in den Diktaten machten als die einsprachigen Kinder, während die anderen paarweisen Vergleiche nicht signifikant wurden. Die Interaktion zwischen beiden Faktoren wurde statistisch ebenfalls nicht signifikant, $F(7.14, 884.71) = 1.54$, $p = .148$, $\eta_p^2 = .012$.

Diskussion

Eine pädagogisch-psychologische Schulleistungsdiagnostik mehrsprachiger Kinder sollte idealerweise auf standardisierte Testverfahren zurückgreifen, deren Fairness für diese Gruppe empirisch geprüft wurde. Denn nur so kann sichergestellt werden, dass die Testergebnisse sinnvoll interpretierbar sind und nicht allein durch die besonderen Spracherwerbsbedingungen zustande kommen. Die Prüfung der Testfairness vieler standardisierter Schulleistungstests im Deutschen steht für mehrsprachige Kinder jedoch noch aus. Daher bestand das erste Ziel dieser Studie darin, für die quantitative Auswertung der Rechtschreibleistung im WÜRT 1 die prozedurale Testfairness für mehrsprachige Erstklässler_innen zu prüfen. Zusätzlich wurde für die qualitative Auswertung des WÜRT 1 überprüft, ob die mehrsprachigen Kinder dieselben Fehlerschwerpunkte in der deutschen Rechtschreibung aufweisen wie die einsprachigen Kinder.

In Bezug auf die erste Fragestellung zeigten lediglich vier der insgesamt 36 Items Hinweise auf DIF – wobei jedoch nur eines dieser Items, das Wort „wünscht“, der Benjamini-Hochberg-Korrektur zur Kontrolle des multiplen Testens standhielt. Das Wort „wünscht“ gehört zu den Mitsprechwörtern und ist laut Testmanual das viertschwerste Wort im gesamten Diktat. Wie die vorliegende Analyse nun zusätzlich verdeutlicht, weist dieses Item für die Gruppe der mehrsprachigen Kinder eine deutlich hö-



Anmerkungen: A = Auslassung eines Lautes, H = Hinzufügen eines Lautes, LV = Lautverwechslung, U = Umstellung von Buchstaben innerhalb eines Wortes, SW = Starke Wortbildveränderung, GK = Verstöße gegen die Groß- und Kleinschreibung, LF = Lautgetreu aber falsch, AW = Auslassen eines ganzen Wortes.

Abbildung 2. Fehlerprofile der einsprachigen Kinder (Linie mit kurzen Strichen), der bilingualen Kinder (durchgezogene Linie) und der Zweitsprachlernenden (Linie mit langen Strichen) im WÜRT 1+ (dargestellt sind die durchschnittlichen Fehleranzahlen der Gruppen).

here Itemschwierigkeit auf als für die einsprachigen Kinder. Dies liegt möglicherweise am Konsonantencluster „nscht", denn ein derart komplexes Konsonantencluster kommt in den anderen Items des WÜRT 1 nicht vor. Dass sich Test*un*fairness bei diesem Item zeigt, ist vor dem Hintergrund einzuordnen, dass fünf aufeinanderfolgende Konsonanten in vielen anderen Sprachen nicht vorkommen, was bei mehrsprachigen Kindern zu Schwierigkeiten bei der phonologischen Differenzierung und entsprechend auch bei der Verschriftlichung führen kann (Dahmen, 2012). Dieses Ergebnis ist des Weiteren interessant, wenn man bedenkt, dass die häufigsten Muttersprachen, die von den mehrsprachigen Kindern dieser Studie gesprochen wurden, Türkisch und Arabisch waren und beide Sprachen über vergleichsweise simple Silbenstrukturen verfügen, bei denen sich im Anlaut höchstens ein und im Auslaut höchstens zwei Konsonanten wiederfinden (siehe Dahmen, 2012).

Zusammengenommen sprechen die vorliegenden Befunde dennoch dafür, den WÜRT 1 für die Diagnostik bei mehrsprachigen Kindern ebenso einzusetzen wie bei einsprachigen Kindern, da nicht davon auszugehen ist, dass die höhere Itemschwierigkeit des Wortes „wünscht" einen erheblichen Einfluss auf das Gesamttestergebnis hat. In der Tat schlagen sich Rohwert-Unterschiede in der Höhe eines einzelnen Punktwertes im WÜRT 1 in Normwert-Unterschieden von lediglich einem (bzw. maximal zwei) T-Wertpunkten nieder. Wird in der vorliegenden Stichprobe beispielsweise das Item „wünscht" bei der Auswertung komplett ausgeschlossen und die neu resultierenden T-Werte beider Gruppen inferenzstatistisch miteinander verglichen, reduziert sich der mittlere Gruppenunterschied unwesentlich um 0.28 T-Wertpunkte. Es kann daher davon ausgegangen werden, dass die Testergebnisse auch bei mehrsprachigen Kindern sinnvoll interpretiert werden können und in gleicher Weise die latente Fähigkeit der Rechtschreibkompetenz widerspiegeln wie bei einsprachigen Kindern. Schwache Testergebnisse sind demnach nicht auf eine systematische Benachteiligung mehrsprachiger Kinder durch das Testverfahren zurückzuführen. Dennoch könnte es für die Individualdiagnostik mitunter sinnvoll sein, bei mehrsprachigen Kindern, die das Item „wünscht" falsch beantwortet haben, zu prüfen, inwiefern eine diagnostische Entscheidung ggf. anders ausfallen würde, wenn das Item richtig beantwortet worden wäre, um auf diese Weise die erhöhte subgruppenspezifische Itemschwierigkeit ins klinische Urteil miteinzubeziehen.

Auch hinsichtlich der zweiten Fragestellung wurden in dieser Studie mehr Gemeinsamkeiten als Unterschiede zwischen den beiden Gruppen deutlich. Zwar machten die mehrsprachigen Kinder – im Einklang mit früheren Befunden (Blatt et al., 2016; May, 2006; Schründer-Lenzen & Merkens, 2006) – insgesamt mehr Fehler in der Rechtschreibung als die einsprachigen Kinder, dennoch zeigten sie vergleichbare Fehlerschwerpunkte. Dass bei den mehrsprachigen Kindern demnach kein qualitativ anderes Fehlerprofil vorlag, legt die Schlussfolgerung nahe, dass eine Rechtschreibförderung an den gleichen Schwerpunkten ansetzen kann wie bei den einsprachigen Kindern. Dennoch muss einschränkend erwähnt werden, dass die Einteilung der Rechtschreibfehler in die vorhandenen Fehlerkategorien des WÜRT 1 lediglich dichotom (d.h. trifft zu, trifft nicht zu) vorgenommen wird. Auch

wird bei den Fehlschreibungen der mehrsprachigen Kinder nicht zusätzlich kodiert, ob es sich dabei um Interferenzfehler handelt oder nicht. Daher erlaubt die vorliegende Studie keine Aussagen darüber, ob und inwiefern sich die Fehlschreibungen *innerhalb* einer Kategorie für die einsprachigen und mehrsprachigen Kinder unterscheiden.

Die häufigsten Fehler betrafen in beiden Gruppen die Groß- und Kleinschreibung, welche jedoch in der 1. Klasse noch nicht systematisch eingeführt ist. Lässt man diese Fehler außer Acht, traten in beiden Gruppen besonders häufig Lautauslassungen und Lautverwechslungen auf. Beides sind Fehlerkategorien, die sich auf Probleme in der phonologischen Bewusstheit und / oder in der Buchstabe-Laut-Verbindung zurückführen lassen (Trolldenier, 2014) und sich durch ein entsprechendes Training beheben lassen. Wenngleich die vorliegende Studie keine Aussage über die Effektivität von Fördermaßnahmen zur phonologischen Bewusstheit und zur Buchstabe-Laut-Verbindung für mehrsprachige Kinder erlaubt, so konnte dies in vorherigen Studien gut belegt werden (Schöppe et al., 2013; Konerding, Bergström, Lachmann & Klatte, 2020). Auch wurden gleichermaßen positive Auswirkung einer solchen Förderung auf die lautgetreue Rechtschreibung bei ein- und mehrsprachigen Kindern gefunden (Weber, Marx & Schneider, 2007).

Ebenfalls nennenswert ist, dass die Kinder dieser Studie besonders viele Fehler in der Kategorie *Starke Wortbildveränderung* zeigten, was sich auch in einer schwachen Gesamtleistung der Stichprobe widerspiegelte und dafür spricht, dass einige Kinder die diktierten Wörter noch sehr fragmentarisch verschriftlichten. Dies ist durchaus bedenkenswert, spricht es doch dafür, dass das grundlegende Prinzip des lauttreuen Schreibens auch am Ende der ersten Klasse von diesen Kindern noch nicht ausreichend verinnerlicht wurde.

Schließlich sollen einige Limitationen der Studie nicht unerwähnt bleiben: Obwohl die vorliegende Stichprobe in einigen Aspekten durchaus vergleichbar ist mit der Normierungsstichprobe des WÜRT 1 (z. B. in Bezug auf die interne Konsistenz oder die Fehlerverteilung), weicht sie in Bezug auf das mittlere Leistungsniveau nach unten ab. Möglicherweise haben sich v. a. Schulen mit einem eher niedrigen Leistungsniveau für die Teilnahme an der Studie angemeldet. Alternativ könnte auch die gewählte Lernmethode im Rechtschreibunterricht eine Rolle spielen. So gehört Hessen (zumindest noch zum Zeitpunkt der Datenerhebung in den Schuljahren 2017/18 und 2018/19) zu den Bundesländern ohne verbindlichen, im Lehrplan verankerten Grundwortschatz, während die Items des WÜRT auf Basis der Grundschulwortschätze von sieben Bundesländern ausgewählt wurden. Dies könnte zu einem Leistungsnachteil für die hessischen Kinder geführt haben. Eine weitere Limitation ist, dass die Gruppe der mehrsprachigen Kinder in der vorliegenden Studie sowohl Kinder umfasste, die Deutsch als Zweitsprache erwarben, als auch Kinder, die Deutsch als eine von mehreren Muttersprachen lernten. Das Zusammenlegen beider Subgruppen war nötig, um eine hinreichend große Stichprobe für die statistischen Analysen zu erhalten. Dennoch sollte bedacht werden, dass sich die beiden Gruppen hinsichtlich ihrer Spracherwerbsbedingungen voneinander unterscheiden: Durch den späteren Erwerbsbeginn und die kürzere Kontaktdauer zur deutschen Sprache zeigen Zweitsprachlernende typischerweise etwas schwächere Sprachleistungen im Deutschen als simultan bilingual aufwachsende Kinder. So fanden beispielsweise Grimm und Schulz (2016), dass simultan bilinguale Kinder in früh erworbenen sprachlichen Phänomenen monolingualen Kindern ähneln, während sie in spät erworbenen Phänomenen Kindern mit Deutsch als Zweitsprache ähneln. Explorativ wurden die varianzanalytischen Analysen daher anschließend zusätzlich mit einer Einteilung in drei Gruppen (einsprachig – zweitsprachlernend – bilingual) durchgeführt. Dabei zeigte sich, dass lediglich die Leistungsunterschiede zwischen den einsprachigen Kindern und denen der Zweitsprachlernenden statistisch signifikant wurden, während die Leistungen der bilingualen Kinder deskriptiv zwischen denen der anderen beiden Gruppen lagen und in keine der beiden Richtungen statistische Signifikanz erreichten. Vor diesem Hintergrund ist nicht auszuschließen, dass sich mehr Items als problematisch hinsichtlich ihrer Testfairness im Vergleich zu einsprachig deutsch aufwachsenden Kindern herausgestellt hätten, wenn nur Zweitsprachlernende in die Analysen einbezogen worden wären. Eine Replikation der vorliegenden Befunde an einer Stichprobe, die ausschließlich Zweitsprachlernende beinhaltet, wäre daher sinnvoll. Ebenso sind Zweitsprachlernende in sich keine homogene Gruppe: Beispielsweise haben Zweitsprachlernende, die schon einen deutschen Kindergarten besuchten, bis zu ihrer Einschulung bereits mehr und intensivere Deutsch sprechende Erwachsenenmodelle erlebt als jene Zweitsprachlernende, die erst ab dem Schulbeginn systematisch Kontakt zur deutschen Sprache haben. Aufgrund des Längsschnittdesigns der vorliegenden Studie wurden jedoch nur Kinder mit Kindergartenbesuch rekrutiert, sodass unklar bleibt, inwiefern sich die hier vorliegenden Befunde zur Testfairness auch auf Zweitsprachlernende ohne Kindergartenbesuch in Deutschland generalisieren lassen.

Diese Limitationen stellen jedoch nicht das Ergebnis in Frage, dass beim Rechtschreibtest WÜRT 1 kein nennenswerter Hinweis auf eine Verletzung der prozeduralen Testfairness bei mehrsprachigen Kindern vorliegt. So zeigte sich nur bei einem einzigen Item ein bedeutsamer Hinweis auf DIF. Dies ist bemerkenswert, da das Thema

Mehrsprachigkeit bei der Itemauswahl scheinbar nicht gesondert berücksichtigt wurde (im Testmanual finden sich zumindest bei den Erläuterungen zur Itemauswahl nur Hinweise auf die Grundschulwortschätze verschiedener Bundesländer, nicht jedoch auf das Thema Mehrsprachigkeit). Der Einsatz des WÜRT 1 zur Rechtschreibdiagnostik und Frühprognose von Rechtschreibschwierigkeiten bei mehrsprachigen Kindern kann anhand der vorliegenden Ergebnisse daher empfohlen werden.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000319

ESM 1. Mplus-Syntax

ESM 2. Tabelle E1. Ergebnisse zum DIF für die Items aus dem WÜRT-Diktat #1 „Auf der Wiese“

ESM 3. Tabelle E2. Ergebnisse zum DIF für die Items aus dem WÜRT-Diktat #2 „Im Krankenhaus“

ESM 4. Tabelle E3. Ergebnisse zum DIF für die Items aus dem WÜRT-Diktat #3 „Im Park“

ESM 5. Tabelle E4. Ergebnisse zum DIF für die Items aus dem WÜRT-Diktat #4 „Im Freien“

ESM 6. Tabelle E5. Post-hoc Vergleiche für die Fehlerkategorien im Rahmen der Messwiederholungs-ANOVA im Zwei-Gruppen-Fall

ESM 7. Tabelle E6. Post-hoc Vergleiche für die Fehlerkategorien im Rahmen der Messwiederholungs-ANOVA im Drei-Gruppen-Fall

Literatur

Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland 2018 – Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung.* Bielefeld: W. Bertelsmann Verlag.

Benjamini,Y. & Hochberg, Y. (1995). Controlling the false discovery rate – A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Methodological, 57*, 289 – 300.

Blatt, I., Prosch, A. & Lorenz, C. (2016). Erhebung der Rechtschreibkompetenz am Ende der Grundschulzeit: Ausgewählte Ergebnisse aus einer Großpilotstudie im Rahmen des Nationalen Bildungspanels. *Zeitschrift für Grundschulforschung*, *9*, 125 – 138.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115 – 124.

Dahmen, S. (2012) Orthografiefehler bei DaZ-Lernern. Ursachen, Diagnostik und Training. In M. Michalak & M. Kuchenreuther (Hrsg.), *Grundlagen der Sprachdidaktik – Deutsch als Zweitsprache.* Baltmannsweiler: Schneider-Verlag.

Dahmen, S. & Weth, C. (2017). *Phonetik, Phonologie und Schrift.* Paderborn: Schöningh.

Daseking, M., Werpup-Stüwe, L., Wienert, L. M., Menke, B. M., Petermann, F. & Waldmann, H.-C. (2015). Sprachfreie Intelligenzdiagnostik bei Kindern mit Migrationshintergrund. *Kindheit und Entwicklung, 24*, 243 – 251. https://doi.org/10.1026/0942-5403/a000180

García, G. E. & Pearson, P. D. (1994). Assessment and diversity. *Review of Research in Education, 20*, 337 – 392.

Gebhardt, M., Rauch, D., Mang, J., Sälzer, C. & Stanat, P. (2013). Mathematische Kompetenz von Schülerinnen und Schülern mit Zuwanderungshintergrund. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.). *PISA 2012 – Fortschritte und Herausforderungen in Deutschland* (S. 275 – 308). Münster: Waxmann.

Grimm, A. & Schulz, P. (2016). Warum man bei mehrsprachigen Kindern dreimal nach dem Alter fragen sollte: Sprachfähigkeiten simultan-bilingualer Lerner im Vergleich mit monolingualen und frühen Zweitsprachlernern. *Diskurs Kindheits- und Jugendforschung, 1*, 27 – 42.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie und Fragebogenkonstruktion* (S. 144 – 171). Berlin: Springer.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1 – 55. https://doi.org/10.1080/10705519909540118

Konerding, M., Bergström, K., Lachmann, T. & Klatte, M. (2020). Effects of computerized grapho-phonological training on literacy acquisition and vocabulary knowledge in children with an immigrant background learning German as L2. *Journal of Cultural Cognitive Science.* https://doi.org/10.1007/s41809-020-00064-3

Lenhard, W. & Lenhard, A. (2017). *Diagnostik von Lesestörungen mit ELFE II bei Kindern mit Migrationshintergrund.* Dettelbach: Psychometrica.

May, P. (2006). Orthographische Kompetenz und ihre Bedingungen am Ende der vierten Jahrgangsstufe. In W. Bos & M. Pietsch (Hrsg.). *Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (S. 111 – 141). Münster: Waxmann.

Muthén, L. K. & Muthén, B. O. (1998 – 2020). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author

OECD (2018). *International Migration Outlook 2018.* OECD Publishing. Paris. https://doi.org/10.1787/migr_outlook-2018-en

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., Li, T. & Menold, N. (2018). Examining measurement invariance and differential item functioning with discrete latent construct indicators: A note on a multiple testing procedure. *Educational and Psychological Measurement, 78*, 343 – 352. https://doi.org/10.1177/0013164416670984

Ruppert, C. & Hanulíková, A. (im Druck). Die Rechtschreibleistung von ein- und mehrsprachigen SchülerInnen: Fehlerraten und Fehlerarten. In K. Nimz, K. Schmidt, & C. Noack (Hrsg.), *Mehrsprachigkeit und Orthographie. Empirische Studien an der Schnittstelle zwischen Linguistik und Erziehungswissenschaft.* Hohengehren: Schneider Verlag.

Schöppe, D., Blatter, K., Faust, V., Jäger, D., Stanat, P., Artelt, C. et al. (2013). Effekte eines Trainings der phonologischen Bewusstheit bei Vorschulkindern mit unterschiedlichem Sprachhintergrund. *Zeitschrift für Pädagogische Psychologie, 27*, 241 – 254. https://doi.org/10.1024/1010-0652/a000110

Schründer-Lenzen, A. & Merkens, H. (2006). Differenzen schriftsprachlicher Kompetenzentwicklung bei Kindern mit und ohne Migrationshintergrund. In A. Schründer-Lenzen (Hrsg.), *Risiko-*

faktoren kindlicher Entwicklung (S. 15 – 44). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90075-9_1

Schwabe, F. & Gebauer, M. M. (2013). (Test-)Fairness – eine Herausforderung an standardisierte Leistungsdiagnostik. In N. McElvany, M. M. Gebauer, W. Bos & H. G. Holtappels (Hrsg.). *Jahrbuch der Schulentwicklung: Daten. Beispiele und Perspektiven* (Bd. 17, S. 217 – 235). Weinheim: Beltz Juventa.

Schwippert, K., Wendt, H. & Tarelli, I. (2012). Lesekompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 191 – 207). Münster: Waxmann.

Stanat, P. & Christensen, G. (2006). *Schulerfolg von Jugendlichen mit Migrationshintergrund im internationalen Vergleich. Eine Analyse von Voraussetzungen und Erträgen schulischen Lernens im Rahmen von PISA 2003.* Berlin: Bundesministerium für Bildung und Forschung.

Tarelli, I., Schwippert, K. & Stubbe, T. C. (2012). Mathematische und naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, H. Wendt, O. Köller & H. Selter (Hrsg.). *TIMMS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 247 – 267). Münster: Waxmann.

Tiedemann, J. & Billmann-Mahecha, E. (2007). Leseverständnis, Familiensprache und Freizeitsprache–Ergebnisse aus der Hannoverschen Grundschulstudie. *Zeitschrift für Pädagogische Psychologie, 21* (1), 41 – 49. https://doi.org/10.1024/1010-0652.21.1.41

Trolldenier, H.-P. (2014). *Würzburger Rechtschreibtest für 1. und 2. Klassen: Ein Verfahren für Grund- und Förderschüler (WÜRT 1 – 2).* Göttingen: Hogrefe.

Weber, J., Marx, P. & Schneider, W. (2007). Die Prävention von Lese-Rechtschreibschwierigkeiten: bei Kindern mit nichtdeutscher Herkunftssprache durch ein Training der phonologischen Bewusstheit. *Zeitschrift für Pädagogische Psychologie, 21*, 65 – 75.

Historie
Onlineveröffentlichung: 10. 05. 2023

Danksagung
Wir danken Prof. Dr. Marcus Hasselhorn für seine Anregungen zum Manuskript und seiner Beteiligung an der TRIO-Studie. Ebenso danken wir Prof. Dr. Diemut Kucharz sowie Prof. Dr. Petra Schulz jeweils mit ihren Teams von der Goethe-Universität Frankfurt für ihr Mitwirken bei der Realisierung der Studie.

Ethische Richtlinien
Eine schriftliche Einverständniserklärung liegt von allen Erziehungsberechtigten der teilnehmenden Kinder vor. Das Forschungsprojekt ist von der zuständigen Ethikkommission ethisch und rechtlich beraten worden.

Förderung
Die Daten entstammen dem Forschungsprojekt „TRIO – Kooperation zwischen Grundschule und Kindertagesstätte – Alltagsintegrierte sprachliche Bildung und Sprachförderung in Kleingruppen“, welches durch das Bundesministerium für Forschung und Bildung (BMBF) finanziert wurde (Förderkennzeichen: 01JI1604 A/B).
Open Access-Veröffentlichung ermöglicht durch die Technische Universität Dortmund.

Dr. Janin Brandenburg
Fakultät Rehabilitationswissenschaften
Fachgebiet Partizipation bei Beeinträchtigungen des Lernens
Technische Universität Dortmund
Otto-Hahn-Straße 6
44227 Dortmund
Deutschland
janin.brandenburg@tu-dortmund.de